DOI: 10.1534/genetics.105.053249

A Novel Method for Detecting Intramolecular Coevolution: Adding a Further Dimension to Selective Constraints Analyses

Mario A. Fares[1] and Simon A. A. Travers

Molecular Evolution and Bioinformatics Laboratory, Department of Biology, National University of Ireland, Maynooth, Ireland

Manuscript received November 8, 2005
Accepted for publication March 3, 2006

ABSTRACT

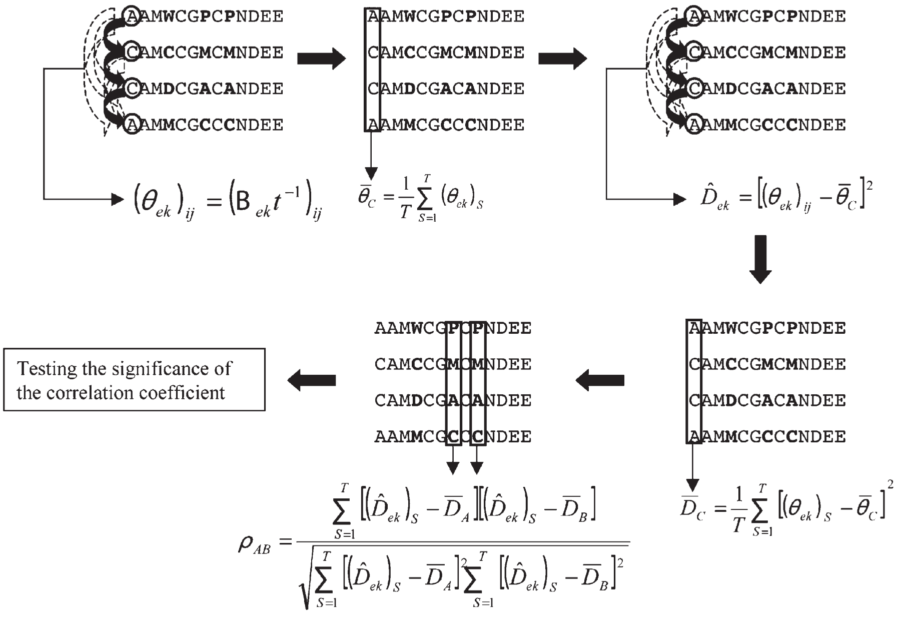
Protein evolution depends on intramolecular coevolutionary networks whose complexity is proportional to the underlying functional and structural interactions among sites. Here we present a novel approach that vastly improves the sensitivity of previous methods for detecting coevolution through a weighted comparison of divergence between amino acid sites. The analysis of the HIV-1 Gag protein detected convergent adaptive coevolutionary events responsible for the selective variability emerging between subtypes. Coevolution analysis and functional data for heat-shock proteins, Hsp90 and GroEL, highlight that almost all detected coevolving sites are functionally or structurally important. The results support previous suggestions pinpointing the complex interdomain functional interactions within these proteins and we propose new amino acid sites as important for interdomain functional communication. Three-dimensional information sheds light on the functional and structural constraints governing the coevolution between sites. Our covariation analyses propose two types of coevolving sites in agreement with previous reports: pairs of sites spatially proximal, where compensatory mutations could maintain the local structure stability, and clusters of distant sites located in functional domains, suggesting a functional dependency between them. All sites detected under adaptive evolution in these proteins belong to coevolution groups, further underlining the importance of testing for coevolution in selective constraints analyses.

UNVEILING the mechanisms of natural selection whereby proteins evolve is one of the fundamental aims of evolutionary genetics studies. The identification of genes showing particular amino acid residues that have undergone adaptive evolution is key in determining functionally or structurally important protein regions. In light of the neutral theory of molecular evolution, mutations are fixed neutrally in proteins (Kimura 1983). Due to their stochastic distribution, only few mutations are beneficial for the biological fitness of the organism and are hence fixed by positive selection (adaptive evolution). However, it is becoming increasingly evident that a significant percentage of genes have undergone adaptive evolution at some stage during their evolutionary past. This body of observed positively selected genes has been growing rapidly during the past decade due to the increase in the number and sensitivity of statistical methods for detecting adaptive evolution.

Methods designed to detect adaptive evolution can be based on Bayesian approaches (Yang *et al.* 2000) or maximum parsimony (Suzuki and Gojobori 1999; Fares *et al.* 2002a). None of these methods takes into account the evolutionary interdependence between protein residues. A protein's function is, however, the result of the functional and structural communication between sites. Sites constraints are hence dependent on the interactions with other residues of the molecule. Mutations at either nearby sites or functionally related distant sites in the structure will change the selective constraints. The more complex the coevolution network is for a particular site, the greater the selection coefficient may be against a mutation at that site due to the dramatic effect that this mutation would have on other coevolving protein regions. Testing coevolution between sites is hence an essential step to complement molecular selection analyses, providing more biologically realistic results.

Grouping sites for molecular evolution analyses has been previously attempted (Hughes and Nei 1988; Clark and Kao 1991). Significant progress has been achieved in building more realistic models (Fares *et al.* 2002a; Suzuki 2004; Berglund *et al.* 2005), albeit several problems regarding the molecular evolution of proteins are still unresolved. For instance, linear sliding-window methods are one-dimensional based and assume independence between different window regions irrespective of their three-dimensional proximity. Conversely, classification of amino acids in the same group of evolution based on their three-dimensional proximity (three-dimensional sliding window) will ignore the coevolution between functional regions that are

[1] *Corresponding author:* Molecular Evolution and Bioinformatics Laboratory, Department of Biology, National University of Ireland, Maynooth, Ireland. E-mail: mario.fares@nuim.ie



$$(\theta_{ek})_{ij} = (B_{ek}t^{-1})_{ij}$$

$$\bar{\theta}_C = \frac{1}{T}\sum_{S=1}^{T}(\theta_{ek})_S$$

$$\hat{D}_{ek} = [(\theta_{ek})_{ij} - \bar{\theta}_C]^2$$

$$\bar{D}_C = \frac{1}{T}\sum_{S=1}^{T}[(\theta_{ek})_S - \bar{\theta}_C]^2$$

$$\rho_{AB} = \frac{\sum_{S=1}^{T}[(\hat{D}_{ek})_S - \bar{D}_A][(\hat{D}_{ek})_S - \bar{D}_B]}{\sqrt{\sum_{S=1}^{T}[(\hat{D}_{ek})_S - \bar{D}_A]^2 \sum_{S=1}^{T}[(\hat{D}_{ek})_S - \bar{D}_B]^2}}$$

Figure 1.—Mathematical framework of the method for detecting coevolution using protein sequences. *B*, *T*, *t*, *D*, ρ, and θ are defined in materials and methods (*Theory*).

spatially distant. Various reports state that residues can form a physically connected network that links distant functional sites in the tertiary protein structure (Süel *et al.* 2003). In fact, mapping energetic interactions in the PDZ domain family predicts a set of energetically coupled positions for a binding site residue that includes unexpected long-range interactions (Lockless and Ranganathan 1999). Coevolution between clusters of sites, which are not in contact, has also been shown (Pritchard and Dufton 2000). Coevolution between distant sites has been observed in sites proximal to regions with critical functions, where coevolution occurs to maintain the structural characteristics around these regions and consequently to maintain the protein conformational and functional stability (Gloor *et al.* 2005).

Coevolution of any type has its origin in the covarion hypothesis proposed first by Fitch and Markowitz (1970). This hypothesis states that, at any given time, some sites are invariable due to their functional or structural constraints but, as mutations are fixed elsewhere in the sequence, these constraints may change. Various methods for identifying covariant amino acid pairs at the molecular level have been previously developed (*e.g.*, Korber *et al.* 1993; Göbel *et al.* 1994; Shindyalov *et al.* 1994; Taylor and Hatrick 1994; Tillier and Collins 1995; Chelvanayagam *et al.* 1997; Pollock and Taylor 1997; Lockhart *et al.* 1998; Tuffley and Steel 1998; Pollock *et al.* 1999; Pritchard *et al.* 2001; Tillier and Lui 2003; Ané *et al.* 2004; Galtier 2004; Dutheil *et al.* 2005). The main limitation of many of these methods has been their inability to separate phylogenetic linkage from functional and structural coevolution. In other methods there was not an assessment of the random noise caused by a limited number of sequences in the alignment or by high pairwise distance. Tillier and Lui (2003) attempted to remove the phylogenetic coevolution; however, their analysis was biased toward those positions covarying with at most a few others. Gloor *et al.* (2005) partially corrected these effects although their method requires alignments of at least 125 sequences to remove stochastic covariation. Further, these methods do not simultaneously take into account the replacement propensity of a site, the background sequence divergence, and the three-dimensional information.

In this work we first develop a novel method for detecting coevolution between sites that allows for more realistic analyses of selective constraints. We then test previously hypothesized and experimentally supported interdomain coevolution in the 90-kDa heat-shock protein (Hsp90), the multimeric Hsp60 (GroEL), and the Gag protein from the human immunodeficiency virus type 1 (HIV-1). We finally apply the new method to uncover coevolution between sites previously detected as under adaptive evolution in GroEL and Gag. We show that coevolution analyses not only provide more realistic results but also highlight the molecular and structural mechanisms shaping the evolution of proteins.

MATERIALS AND METHODS

Theory: Coevolution analysis using protein sequences (CAPS) compares the correlated variance of the evolutionary rates at two sites corrected by the time since the divergence of the protein sequences they belong to (Figure 1). Substitutions or conservation at two independent sites cannot be directly compared due to their amino acid composition difference. The method instead compares the transition probability scores between two sequences at these particular sites, using the blocks substitution matrix (BLOSUM) (Henikoff and Henikoff 1992). For each protein alignment the correspondent BLOSUM matrix is applied, depending on the average sequence identity.

Despite the fact that BLOSUM matrices correct for the substitution values due to the estimated divergence between sequence pairs, a given alignment can include sequences whose pairwise distance is significantly divergent from the mean pairwise distance. For instance, an alignment including two highly divergent sequence groups (for example, gene duplication predating speciation) could show an unrealistic pairwise average identity level. In this respect, sequences that diverged a long time ago are more likely to fix correlated mutations at two sites by chance (under a Poisson model) compared to recently diverged sequences. BLOSUM values should be hence normalized by the time of divergence between sequences. BLOSUM values (B_{ek}) are thus weighted for the transition between amino acids e and k using the time (t) since the divergence between sequences i and j:

$$(\theta_{ek})_{ij} = (\mathrm{B}_{ek}t^{-1})_{ij}. \tag{1}$$

The assumption made in Equation 1 is that the different types of amino acid transitions (slight to radical amino acid changes) in a particular site follow a Poisson distribution along time. The greater the time is since the divergence between sequences i and j the greater the probability is of having a radical change. A linear relationship is thus assumed between the BLOSUM values and time. Synonymous substitutions per site ($d_{\mathrm{S}_{ij}}$) are silent mutations, as they do not affect the amino acid composition of the protein. These mutations are therefore neutrally fixed in the gene. Assuming that synonymous sites are not saturated or under constraints, d_{S} is proportional to the time since the two sequences compared diverged. Time (t) therefore is measured as d_{S}. Note that convergent radical amino acid changes at two sites in sequences that have diverged recently have larger weights compared to convergent changes in distantly related sequences.

The next step is the estimation of the mean θ-parameter for each site ($\bar{\theta}_{\mathrm{C}}$) of the alignment, so that

$$\bar{\theta}_{\mathrm{C}} = \frac{1}{T}\sum_{S=1}^{T}(\theta_{ek})_S. \tag{2}$$

Here S refers to each pairwise comparison, while T stands for the total number of pairwise sequence comparisons, and thus

$$T = \frac{N(N-1)}{2}, \tag{3}$$

where N is the total number of sequences in the alignment.

The variability of each pairwise amino acid transition compared to that of the site column is estimated as

$$\hat{D}_{ek} = [(\theta_{ek})_{ij} - \bar{\theta}_{\mathrm{C}}]^2. \tag{4}$$

The mean variability for the corrected BLOSUM transition values is

$$\bar{D}_{\mathrm{C}} = \frac{1}{T}\sum_{S=1}^{T}[(\theta_{ek})_S - \bar{\theta}_{\mathrm{C}}]^2. \tag{5}$$

The coevolution between amino acid sites (A and B) is estimated thereafter by measuring the correlation in the pairwise amino acid variability, relative to the mean pairwise variability per site, between them. We thus use the relative variability rather than the absolute variability to measure the correlation between two sites. This ensures making the covariation independent from the differences in the rates of evolution of the sites compared. This covariation is measured as the correlation between their $\hat{D}_{ek}$-values, such as

$$\rho_{AB} = \frac{\sum_{S=1}^{T}[(\hat{D}_{ek})_S - \bar{D}_A][(\hat{D}_{ek})_S - \bar{D}_B]}{\sqrt{\sum_{S=1}^{T}[(\hat{D}_{ek})_S - \bar{D}_A]^2\sum_{S=1}^{T}[(\hat{D}_{ek})_S - \bar{D}_B]^2}}. \tag{6}$$

Here e and k are any two characters at sites A and B. To determine if the correlation coefficient (ρ_{AB}) is significant, either a resampling or a simulation analysis can be performed. In the first approach we randomly sample K numbers of pairs of sites and compute Equations 1–6 for each pair. The mean correlation coefficient and its variance are then estimated as

$$\bar{\rho} = \frac{1}{K}\sum_{l=1}^{K}\rho_l \quad \text{and} \quad V(\rho) = \frac{1}{K}\sum_{l=1}^{K}(\rho_l - \bar{\rho})^2. \tag{7}$$

Correlation coefficients are then tested for significance under a normal distribution:

$$Z = \frac{\rho_{AB} - \bar{\rho}}{\sqrt{V(\rho)}}. \tag{8}$$

The second approach consists of the Monte Carlo simulation of K sequence alignments. Here the coevolution test is conducted for a number of randomly selected pairs of sites in each simulated alignment or for the complete set of pairs of sites in the random data set computing Equations 1–6. An average value of the correlation for the simulated alignments and its variance are obtained utilizing Equation 7. Finally, the real correlation coefficients are tested using Equation 8.

The statistical power of the test is optimized by analyzing sites showing

$$\bar{D}_{\mathrm{C}} > \Theta - 2\sigma_{\Theta}. \tag{9}$$

Here, Θ is the parametric value of $\bar{D}_{\mathrm{C}}$ from Equation 5 and σ is the standard deviation of Θ. Θ is calculated as

$$\Theta = \frac{1}{L}\sum_{s=1}^{L}(\bar{D}_{\mathrm{C}})_s, \tag{10}$$

where L is the length of the alignment. Pairwise comparisons including gaps in any or both sites at any sequence are excluded from the analysis.

Removing the phylogenetic coevolution: Coevolution between amino acid sites can be the result of their structural, functional, or physical interaction; their phylogenetic convergence; and their stochastic covariation. The analysis of simulated data to test for significance removes stochastic effects. To disentangle functional, structural, and interaction coevolution from phylogenetic coevolution, the method is applied to the complete alignment and to subalignments, where specific phylogenetic clades are removed from the tree. Coevolving amino acid sites that are no longer detected following removal of one of the clades will be classified as phylogenetic coevolving sites as they occur in specific branches of the tree. Conversely, coevolving amino acid sites detected irrespective of the tree clades removed will be considered as functional/structural/interaction coevolving sites since they present correlated changes throughout the phylogenetic tree. Note that the latter condition means that when one amino acid changes, the covarying amino acid has necessarily to change. In the former condition, a change in one site does not always (in all branches) involve a change in the covarying site. In other words, our method detects phylogenetic-independent coevolution. Clades for coevolution analyses are defined in terms of their biological coherence and/or statistical support (defined as bootstrap values). Consequently, phylogenetic clades are specified prior to conducting the coevolutionary analysis and they include sequences that are forming either a well-defined biological cluster or alternatively a cluster supported by a high bootstrap value.

Using the atomic distances as additional information in coevolution analyses: Spatial proximity between coevolving sites can be used to define their structural or functional interaction. In this method coevolution is not always synonymous with physical interaction but also involves structural and functional coevolution, as has been previously described (Lockless and Ranganathan 1999; Pritchard and Dufton 2000; Süel *et al.* 2003; Gloor *et al.* 2005).

The three-dimensional closeness of two sites is estimated as the vectorial distance between their atomic centers (δ). This distance is obtained by comparing the three-dimensional coordinates (X, Y, and Z) of atoms A and B for amino acids i and j:

$$\delta_{A-B} = \vec{A} - \vec{B} = \sqrt{(X_A - X_B)^2 + (Y_A - Y_B)^2 + (Z_A - Z_B)^2}. \tag{11}$$

Since each amino acid consists of several atoms, the mean atomic distance ($\bar{\delta}$) between sites i and j is taken:

$$\bar{\delta}_{i-j} = \sqrt{\begin{aligned} &\left[\left(\frac{1}{\mu_i}\sum_{m=1}^{\mu_i} X_m\right)_i - \left(\frac{1}{\mu_j}\sum_{m=1}^{\mu_j} X_m\right)_j\right]^2 \\ &+ \left[\left(\frac{1}{\mu_i}\sum_{m=1}^{\mu_i} Y_m\right)_i - \left(\frac{1}{\mu_j}\sum_{m=1}^{\mu_j} Y_m\right)_j\right]^2 \\ &+ \left[\left(\frac{1}{\mu_i}\sum_{m=1}^{\mu_i} Z_m\right)_i - \left(\frac{1}{\mu_j}\sum_{m=1}^{\mu_j} Z_{mj}\right)_j\right]^2 \end{aligned}}. \tag{12}$$

Here, μ refers to the total number of atoms in the amino acid. The significance of the distance is tested by comparing it to a distribution of K random amino acid pairs sampled from the three-dimensional structure. The reason for conducting a statistical analysis to detect proximal sites is that sites may be considered significantly proximal or distant depending on the shape of the protein. On the other hand, sites that are not in physical contact but are surrounding functionally important sites, and are hence proximal, can present coevolution due to their proximity to important sites (Gloor *et al.* 2005).

Simulation studies: We tested the sensitivity of CAPS using simulated data that allow for the control of the extent of coevolution and the evolutionary history of each site in the alignment. We also compared CAPS with other nonparametric methods that use the information theory or a Bayesian approximation, including the method of Korber [herein called the mutual information criterion (MICK) implemented in our program PIMIC and available on request; Korber *et al.* 1993] and the method of Tillier and Lui (2003) implemented in the program Dependency, as well as with the parametric method of Pollock *et al.* (1999) implemented in the program lnLCorr. While parametric methods can be more powerful than nonparametric ones, incorrect assumptions in the model can yield a high number of false positives (Dimmic and Hubisz 2005). We thus compared CAPS to more conservative nonparametric methods and to more powerful parametric methods. We simulated sequence alignments using a model similar to that devised by Pollock *et al.* (1999). Initially an ancestral sequence of 200 amino acids was generated using the amino acid composition corresponding to the equilibrium residue frequencies in naturally occurring proteins (Jones *et al.* 1992). This sequence was evolved using a Markov chain Monte Carlo simulation along a phylogenetic tree and, simultaneously, 10 pairs of sites were randomly selected to coevolve. Coevolution was established by forcing correlated variation such as the transition between amino acids at both sites had similar θ_{ek} as specified in Equation 1. We have also introduced coevolution without forcing this condition but results were unaltered. The phylogenetic trees used for simulations were bifurcated and symmetric (all branches had the same length). The robustness of the coevolution analysis to the sequence phylogenetic divergence was assessed using different levels of background noise and alignment sizes. We fixed the background noise (branch lengths) to evolve the ancestral sequence at 0.1, 0.2, 0.5, and 1 Poisson-corrected substitutions per site. The numbers of sequences tested were 10, 20, and 30 sequences. We simulated and tested multiple simulations for each condition (number of sequences and number of substitutions per site), performing 12 different types of simulation analyses. We measured the sensitivity (SN) of the method as

$$\mathrm{SN} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}. \tag{13}$$

Here TP and FP are the numbers of true and false positives, respectively. We also measured the specificity of the methods as:

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}}. \tag{14}$$

Here, TN and FN are the numbers of true and false negatives, respectively.

Analysis of real sequences: We used CAPS to analyze the Gag protein from HIV-1, the 90-kDa heat-shock protein (Hsp90), and the 60-kDa heat-shock protein (GroEL). These proteins are multimeric and organized in functionally connected domains and are hence perfect for testing interdomain coevolution. To show an example of coevolution between sites that have undergone positive selection we used previously published data demonstrating adaptive evolution in GroEL from endosymbiotic bacteria of insects (Fares *et al.* 2002b, 2004) and the HIV-1 *gag* gene (Yang *et al.* 2003).

Gag protein from HIV-1: The HIV-1 genome is translated to yield both structural and nonstructural proteins (Wain-Hobson *et al.* 1985). Among these proteins, Gag is a 55-kDa polyprotein that is initially associated with the cell membrane to ease the budding of virus particles from the host cell. Gag is further processed to produce four proteins called matrix (p17), capsid (p24), nucleocapsid (p9), and p6 (Gottlinger *et al.* 1989). We tested whether coevolution exists between specific Gag proteins or amino acid sites in specific HIV-1 lineages. The HIV-1 group M subtypes are thought to have originated from a single ancestor and are currently described by highly supported clades.

Hsp90: Hsp90 is an ATPase molecular chaperone that assists the conformational maturation of molecules involved in cell-cycle regulation and signal transduction (Pratt 1998; Buchner 1999; Caplan 1999; Mayer and Bukau 1999). Hsp90 is translated as a monomeric protein but its function depends on its dimerization. Several functional domains can be identified in the linear sequence of Hsp90 (supplemental Table 1 at http://www.genetics.org/supplemental/). The importance of the complex intramolecular interactions for the Hsp90 function is poorly understood (Prodromou *et al.* 1999; Johnson *et al.* 2000; Chadli *et al.* 2000). Here we apply CAPS to test and identify interdomain coevolution within Hsp90.

Analysis of the heat-shock protein 60-kDa GroEL: The ATPase molecular chaperone GroEL is found specifically in bacteria and the organelles of eukaryotic cells (Landry *et al.*

1993). The multimeric protein GroEL folds 10–15% of slow-folding proteins, which are mostly aggregation prone (Deuerling *et al.* 1999; Thulasiraman *et al.* 1999). Each GroEL subunit is organized in three domains; apical, equatorial, and intermediate (Braig *et al.* 1994, 1995). Several functionally important intradomain regions in GroEL have been previously identified (supplemental Table 2 at http://www.genetics.org/supplemental/). Here we test if coevolution among sites is crucial for the functional and structural stability of GroEL.

Sequence alignments and phylogenetic inferences: GenBank accession numbers for the *gag*, *hsp90*, and *groel* sequences are provided in supplemental Tables 3, 4, and 5, respectively, at http://www.genetics.org/supplemental/. Protein sequences were aligned (available from the corresponding author upon request) using CLUSTAL X (Jeanmougin *et al.* 1998). Nucleotide sequences were then aligned, concatenating triplets according to the amino acid sequence alignment.

The phylogeny of the HIV-1 group M subtypes is very well defined (Robertson *et al.* 2000). Representative sequences were selected in the manner described previously (Travers *et al.* 2005). For each subtype, all available full-genome sequences were retrieved from the Los Alamos HIV database (http://hiv-web.lanl.gov) and a neighbor-joining tree of each resulting data set was reconstructed using PAUP* 4.0b10 (Swofford 1998). Representative sequences were selected for each subtype on the basis of their spread throughout the subtype tree, resulting in the selection of a diverse range of sequences for each subtype (supplemental Table 3 at http://www.genetics.org/supplemental/). The resulting data set contained 36 taxa.

For Hsp90 we used sequences from unicellular and multicellular eukaryotes comprising a total of 43 taxa. Aligned sequences were subject to phylogenetic analyses using PAUP* 4.0b10. Maximum-likelihood and maximum-parsimony analyses yield the same phylogenetic tree. We used the GroEL phylogenetic tree obtained in a previous work (Fares *et al.* 2002b).

Analysis of coevolution: Coevolution analyses were implemented in the program CAPSv1.0 (available from the corresponding author on request). Synonymous substitutions d_S were considered to be proportional to the time since the divergence between sequences since no indication of saturation of synonymous sites was observed using SWAPSC v1.0 (Fares 2004). The significance of the correlation coefficients was tested using 10,000 pseudorandom pairs of amino acid sites and a confidence value of ($\alpha = 0.001$), to minimize type I error. Clades defined in each protein for coevolution analyses are indicated in Figures 2, 3A, and 4A.

RESULTS

Testing the accuracy of CAPS: The analysis of simulated data sets demonstrates that CAPS is highly sensitive and robust at a wide range of amino acid distances and alignment sizes (Figure 2, A–C). CAPS sensitivity ranged between 65 and 87% in alignments of 10 sequences (Figure 2, A–C). Increasing the alignment length to 20 and 30 sequences yielded sensitivity values between 80 and 90% and between 83 and 98%, respectively.

Although all methods find a high percentage of true coevolutionary amino acid pairs (Figure 2, D–F), the alternate methods to which CAPS was compared also identified large numbers of false positives, showing sensitivity values ranging between 8 and 17% in MICK, between 7 and 15% in Dependency, and between 7 and 8% in lnLCorr (Figure 2A). When the alignment size increased to 20 sequences, the sensitivity of MICK improved at all the distances, ranging between 20 and 15% (Figure 2B). Dependency and lnLCorr presented lower sensitivity values compared to MICK, ranging between 8 and 10% (Figure 2B). Using alignments containing 30 sequences, the sensitivity of MICK, Dependency, and lnLCorr decreased as the average sequence pairwise distance increased (Figure 2C). MICK, Dependency, and lnLCorr seem to require very dense phylogenetic trees as previously suggested (Pollock *et al.* 1999; Tillier and Lui 2003). Using alignments of 200 sequences does not seem to change the sensitivity of CAPS (data not shown). The sensitivity of CAPS increases with the number of sequences in the alignment (using a multivariate test, $F = 9.968$, $P = 0.016$), unlike the other methods, which present no evidence for such an increase (Figure 3A). Conversely, the level of pairwise amino acid divergence negatively affects all the methods except MICK ($F = 1.446$, $P = 0.309$) (Figure 3B).

We analyzed the effect of two factors, alignment size and amino acid substitutions per site, on the sensitivity of all the methods in general and of each method individually. A multivariate test demonstrates that neither of these factors alone influences the sensitivity of the methods for detecting coevolution ($F = 1.951$ and $F = 2.301$ and $P = 0.267$ and $P = 0.090$, for the effects of the alignment size and amino acid distance, respectively). The interaction of both factors, however, has a significant effect on the sensitivity of the four methods taken together ($F = 131.938$; $P \ll 0.001$) or individually (data not shown). In summary, a low number of sequences combined with a high pairwise sequence divergence negatively affect the sensitivity of all the methods for detecting coevolution.

Analysis of the specificity of the methods demonstrates that CAPS is more specific than the alternate methods, although specificity was detected to be significant in all four methods used (data not shown).

Phylogenetic and functional coevolution in Gag from HIV-1: Following coevolution analysis of the complete *gag* alignment, we identified 21 groups of coevolving sites, containing 73 unique residues (Figure 4A). Of these 73 residues, 42 were observed to have undergone phylogenetic coevolution, in that removal of a particular clade in the analysis resulted in loss of detection of that site in the subsequent analysis using CAPS. In all lineages coevolving sites were spread throughout the gene with the majority of sites present in the functional p17 and p24 regions (Figure 4B). Interestingly, while subtypes A and G, which evolved from a common ancestor, have the highest number of phylogenetically coevolving sites only three residues (E107, I147, and T186) are shared between them, indicating the presence of both lineage-specific and ancestral coevolution within the *gag* gene (Figure 4B).

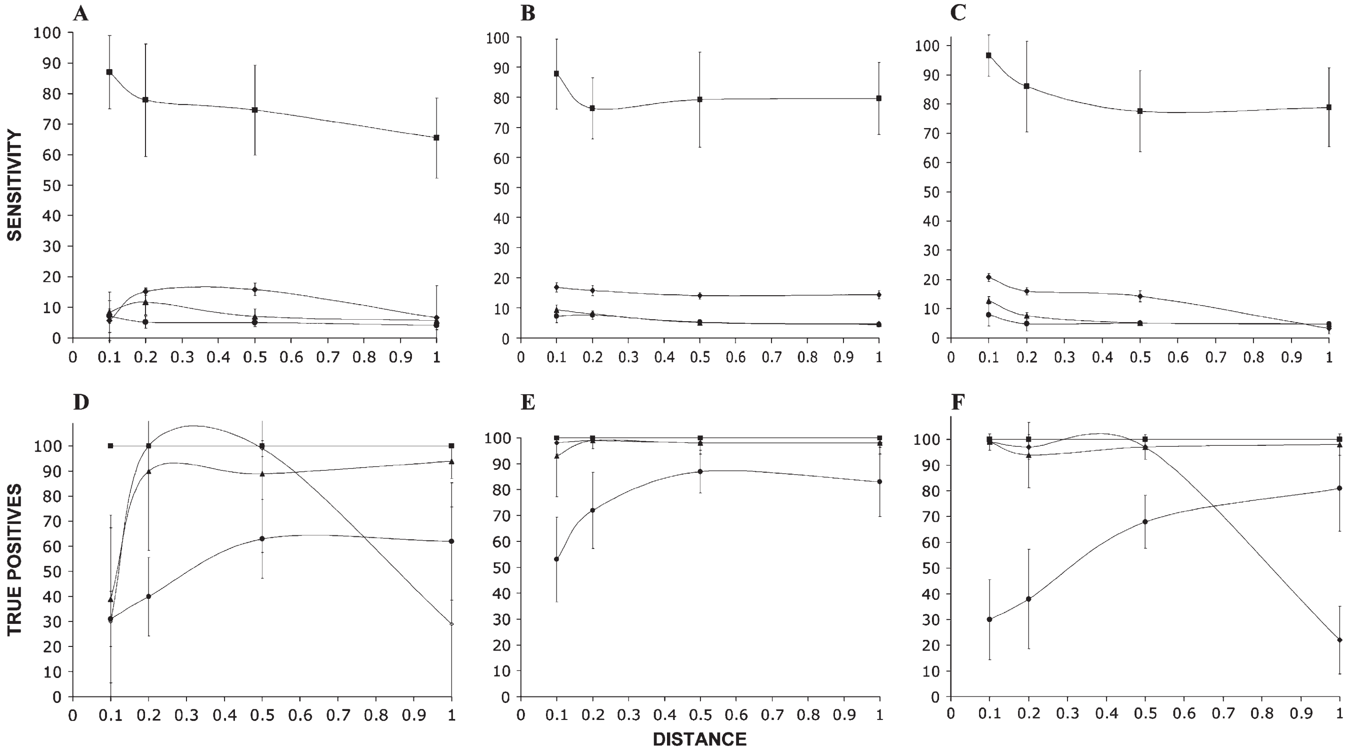


Figure 2.—Plots of the sensitivity (*y*-axis; plots A–C) and the percentage of true positive covariation pairs (*y*-axis, plots D–F) against the Poisson-corrected amino acid distance per site (*x*-axis). Sensitivity and the percentage of true positives have been tested using 10, 20, and 30 sequences as indicated in plots A–F. The sensitivity of CAPS has also been compared to that of the nonparametric mutual information criterion (MICK) of Korber *et al.* (1993) and the method of Tillier and Lui (2003) and to that of the parametric method of Pollock *et al.* (1999). Error bars for the mean sensitivity values over the simulations are also shown. ♦, MICK; ▲, Dependency; ■, CAPS; ●, lnLCorr.

Also, numerous examples of convergent coevolution were observed between subtypes (Figure 4B).

Two groups of amino acid sites (R15, K28, R91, A120 and E55, V128) maintained their grouping no matter what clade (defined as subtype) was removed, indicating functional/structural/physical interaction (herein called functional groups, FGs) coevolution between these sites. We observed 8 of the 14 positively selected amino acids identified by Yang *et al.* (2003) as coevolving within the HIV-1 group M phylogeny (R15, K28, G62, Q69, R91, I138, N252, and T280) with 5 of these (K28, G62, Q69, R91, and T280) having been identified by Yang *et al.* as undergoing adaptive evolution in separate analyses of subtypes A, B, and C. Of the 8 residues overlapping between our study and that of Yang *et al.*, 3 (R15, K28, and R91) are present in one of the two FGs.

Important functional regions in Hsp90 do coevolve: The clades defined for the coevolution analyses were those defined as biologically distinguishable organisms, including Hsp90 from endoplasmic reticulum, chloroplast, yeast, plants, protozoan parasites, insects, and high eukaryotes (Figure 5A). The application of the coevolution analysis to Hsp90 identified 34 groups of coevolution (G1–G34; Figure 5B). A significant proportion of the coevolving sites have an important reported biological function (Table 1 and Figure 5B).

We applied a neural network-based analysis implemented in the program CONSEQ (Berezin *et al.* 2004) to predict the functional or structural importance of residues never tested experimentally. To avoid missing information due to possible shifts in the selective constraints acting on Hsp90 residues in the branch separating unicellular from multicellular eukaryotes, we applied CONSEQ to both data sets (each one comprising a total of 20 sequences). Caution is required when using this approach due to problems in the specificity of the method (Berezin *et al.* 2004). We found that sites reported to be important in the literature were identified by the method. In addition, in most of the cases, sites predicted to be buried or exposed were correctly identified when compared to the three-dimensional structure. After using this approach, and taking into account the literature published, the average number of functionally or structurally important sites in each group of coevolution was 86.26% (Figure 5B). A careful inspection of Figure 5B reveals that coevolution has occurred within and between functional domains. In fact, coevolution was found between domains involved in protein interaction (PI); in ATP modulatory, ATP amino-, and carboxy-terminal binding regions (ATPM, ATP-Nt, and ATP-Ct); and in dimerization domains (DD) (Table 1).

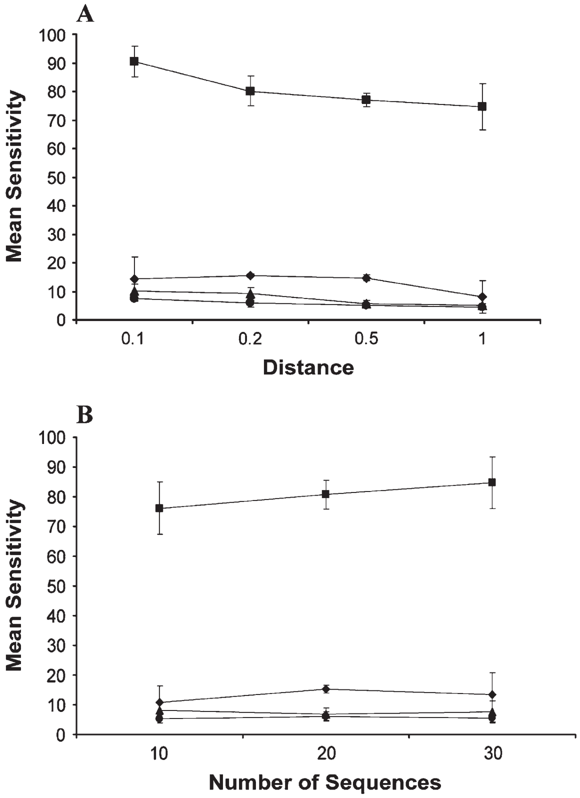


FIGURE 3.—Effect of the alignment size and Poisson-corrected amino acid distance per site on the sensitivity of the methods for detecting coevolution. (A) Plot of the mean sensitivity over the three sets of alignment sizes (10, 20, and 30 sequences) against the Poisson-corrected amino acid distance per site. (B) Plot of the mean sensitivity over the level of Poisson-corrected amino acid distance per site against the alignment size. Error bars for the mean sensitivity values are also shown. ♦, MICK; ▲, Dependency; ■, CAPS; ●, lnLCorr.

The mean variance for the amino acid transition in each group of coevolution ($\bar{D}_C$) ranged between 0.531 and 2.084, whereas the mean correlation ($\bar{\rho}$) varied between 0.625 and 0.948 (Table 1). The three-dimensional structure is available only for the amino-terminal and middle segments. The analysis of the atomic distances (AD) identified a certain percentage of coevolving residues within each group as spatially close (Table 1). Examples are the pairs of amino acids [(S113, A112) and (E4, G3, and V163)] in the amino-terminal domain (Figure 5C) and the groups of amino acids [(E351, N382), (Q409, F410), and (K445, S446)] in the middle segment (Figure 5D). A number of coevolving groups exhibit coevolution between sites significantly distant but functionally related (Table 1).

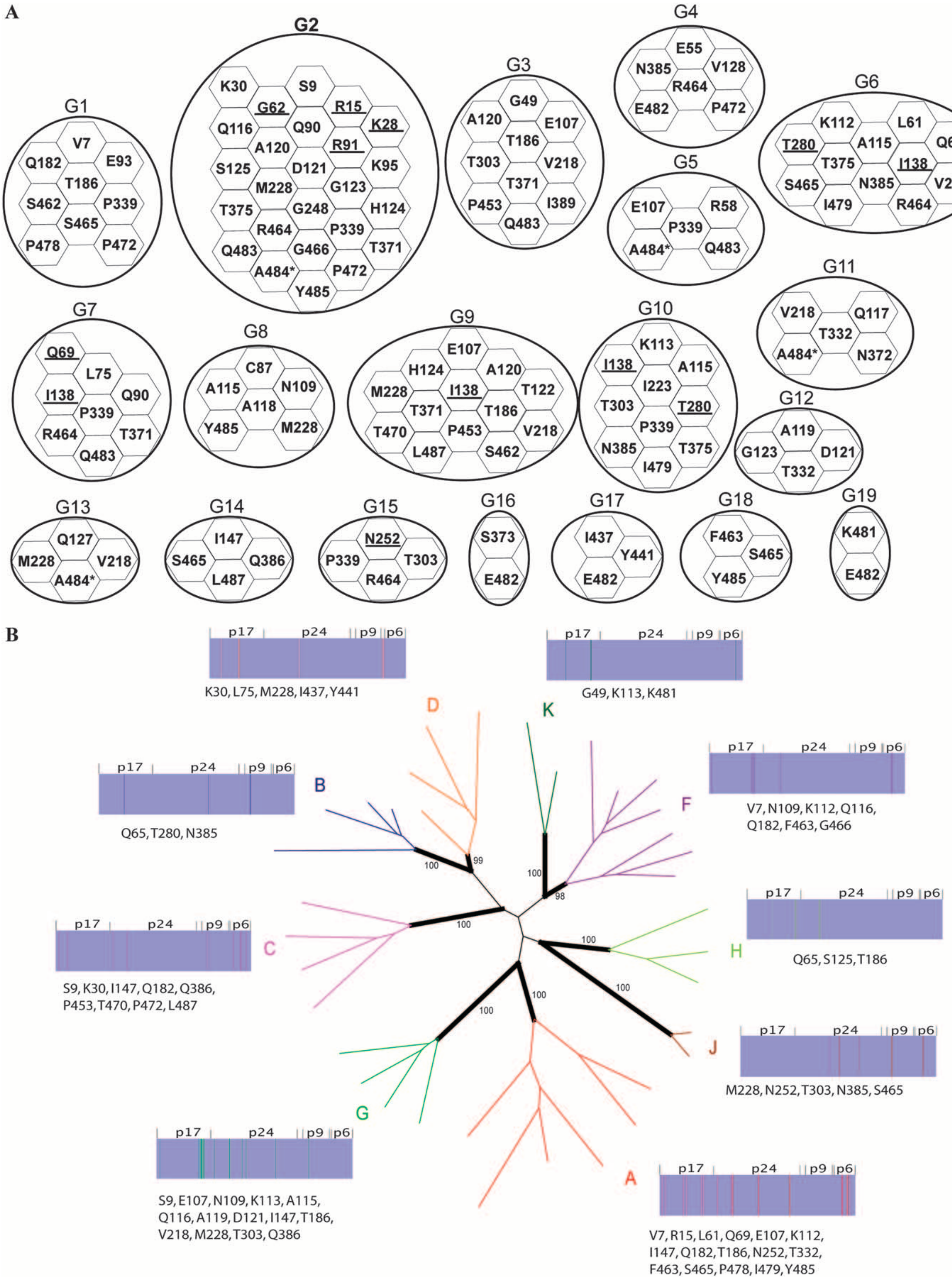
Revealing adaptive coevolution in GroEL: Application of CAPS to GroEL, taking into account the clusters belonging to different bacterial groups and defined in Figure 6A, identified 17 groups of coevolving amino acids (G1–G17; Figure 6B). Seventy-five percent of coevolving sites were previously reported in functional data (supplemental Table 2 at http://www.genetics.org/supplemental/), or by neural network analysis, as functionally or structurally important.

The mean variance for the amino acid transition in each group of coevolution ($\bar{D}_C$) ranged between 0.739 and 2.477, while the mean correlation ($\bar{\rho}$) varied between 0.510 and 0.944 (Table 2). All of the groups detected included sites belonging to the apical and equatorial domains with very few sites belonging to the intermediate domain. Most of the sites not identified by functional data or undetected by predicting neural networks presented significantly short distances to functionally or structurally important sites. Examples of spatially close sites were sites [(K132, L134, K425), (A127, K132), and (S424, K425)] in the equatorial domain and sites (T210, G211, V213) in the apical domain (Figure 6C).

Interestingly, a significant proportion of the sites detected as coevolving in this GroEL data set have been previously proposed to be under adaptive evolution (Figure 6B) (FARES *et al.* 2002b, 2004). Most of these positively selected sites included in the same coevolution group belong to different domains (apical and equatorial domains). Examples of this were found in the coevolution groups G1, G2, G7, G8, and G9 (Figure 6B). Taking into account all these data, the overall percentage of coevolving sites previously identified as key for GroEL function or evolution was 82%. Other positively selected sites were proximal to important sites (for example, sites T210, G211, and V213 in group G1 are in physical contact with sites S201, Y203, and F204 that are involved in GroES and substrate binding; supplemental Table 2 at http://www.genetics.org/supplemental/). An important observation to mention is that not all positively selected sites in GroEL were detected to be under coevolution, which makes the dependence between positive selection and coevolution less likely.

DISCUSSION

A highly sensitive method: To assess the validity of the method for detecting coevolution between sites, we asked two questions: How sensitive is the method to distinguish between true and false positives? And, how much does the method improve upon the sensitivity of similar nonparametric and parametric methods previously published? Our simulations indicate that CAPS does indeed identify a high percentage of true correlated pairs even at large pairwise sequence distances. When considering sensitivity, CAPS performs significantly better than the other three methods to which it was compared over all distances and sequence numbers. We also followed the recommendations of TILLIER and LUI (2003), increasing the number of sequences in our simulated alignments to a number equivalent to the number of amino acids sites, but the sensitivity of CAPS remained unaltered (data not shown). The importance of the number of sequences in the alignment


A
G1
V7
Q182
E93
T186
S462
P339
S465
P478
P472
G2
K30
S9
G62
R15
Q116
Q90
K28
A120
R91
S125
D121
K95
M228
G123
T375
G248
H124
R464
P339
Q483
G466
T371
A484*
P472
Y485
G3
G49
A120
E107
T186
T303
V218
T371
P453
I389
Q483
G4
E55
N385
V128
R464
E482
P472
G5
E107
R58
P339
A484*
Q483
G6
K112
L61
T280
A115
Q65
T375
I138
S465
N385
V218
I479
R464
G7
Q69
L75
I138
Q90
P339
R464
T371
Q483
G8
C87
A115
N109
A118
Y485
M228
G9
E107
H124
A120
M228
I138
T122
T371
T186
T470
P453
V218
L487
S462
G10
K113
I138
A115
I223
T303
T280
P339
N385
T375
I479
G11
V218
Q117
T332
A484*
N372
G12
A119
G123
D121
T332
G13
Q127
M228
V218
A484*
G14
I147
S465
Q386
L487
G15
N252
P339
T303
R464
G16
S373
E482
G17
I437
Y441
E482
G18
F463
S465
Y485
G19
K481
E482
B
p17 p24 p9 p6
K30, L75, M228, I437, Y441
p17 p24 p9 p6
G49, K113, K481
D
K
p17 p24 p9 p6
B
Q65, T280, N385
p17 p24 p9 p6
F
V7, N109, K112, Q116, Q182, F463, G466
99
100
100
98
p17 p24 p9 p6
C
S9, K30, I147, Q182, Q386, P453, T470, P472, L487
100
100
p17 p24 p9 p6
H
Q65, S125, T186
100
100
100
p17 p24 p9 p6
J
M228, N252, T303, N385, S465
G
p17 p24 p9 p6
S9, E107, N109, K113, A115, Q116, A119, D121, I147, T186, V218, M228, T303, Q386
A
p17 p24 p9 p6
V7, R15, L61, Q69, E107, K112, I147, Q182, T186, N252, T332, F463, S465, P478, I479, Y485

to obtain accurate coevolution results, especially in mutual information-based methods, has been previously investigated by Gloor *et al.* (2005). They reported that these methods require alignment sizes 10 times greater than those used in this study. We have shown that CAPS exhibits high sensitivity, using either large or small data sets. Our method therefore has an enormous advantage in the analysis of proteins that have not been sequenced in many organisms, proteins that arose recently in evolution, or proteins in which fast evolution permits obtaining accurate results only at very narrow phylogenetic ranges.

Coevolution within the *gag* gene phylogeny: We have identified coevolution between the different functional regions of the *gag* gene in HIV-1 (Figure 3A). The two groups of amino acid sites exhibiting coevolution regardless of clade removal (R15, K28, R91, A120 and E55, V128) are detected as having a functional or structural dependency and included sites previously detected as having undergone adaptive evolution. Also, 42 of 73 sites were detected as coevolving phylogenetically with most of these coevolving sites being unique to specific lineages, thereby providing evidence for possible selective shifts between subtypes within the HIV-1 group M phylogeny. Interestingly, some of those sites were identified as phylogenetic ancestral coevolution (coevolving residues in the branch leading to the ancestor of two subtypes) or convergent coevolution (between lineages that do not share a recent common ancestor; Figure 4B). These results provide further evidence of the need for a more comprehensive evolutionary analysis of the distinct HIV-1 group M subtypes to obtain a full understanding of past, present, and future HIV-1 dynamical change.

Coevolution between functional domains in Hsp90: Several studies demonstrate that Hsp90 functions through marked conformational changes that are governed by complex interdomain interactions (*e.g.*, Prodromou *et al.* 1999; Chadli *et al.* 2000). Because of this complexity, the number of experimental analyses required to fully understand intramolecular Hsp90 interactions is prohibitively high. Moreover, experimental identification of residues involved in the stability of interdomain interactions due to their spatial proximity to functionally important domains is anything but straightforward. Computational methods are hence instrumental in the detection and *a priori* identification of regions or domains in which functional interaction should be tested.

CAPS identified 100% of the Hsp90 interdomain interactions proposed in previous studies and uncovered potential interactions that should be tested. We detected coevolution between sites belonging to PI2 and PI3 domains that are involved in binding different protein clients (Table 1). Simultaneous binding of proteins by Hsp90 has been proposed by other authors who suggest a synergistic effect of these proteins in the function of Hsp90 (Chen *et al.* 1998). An example of simultaneous substrate binding is that provided by the eNOS activation pathway, where the upstream activator of PKB/Akt, PDK1, binds Hsp90 simultaneously to eNOS (Fujita *et al.* 2002). Coevolution between these domains has interesting implications for the simultaneous and regulated binding of proteins with Hsp90. The method also detects coevolved amino acid sites involved in domain dimerization (DD2 and DD3) and cochaperone binding. Sites belonging to DD2 have been associated with the binding of accessory proteins such as Hop and immunophilines that are essential for the interaction and maturation of the complex Hsp90-Hsc70-client proteins (Chen *et al.* 1998). Deletion of the region 661–677 from the Hsp90 of the chicken *Gallus gallus* results in loss of Hsp90 dimerization and diminished interactions with all cofactors (Chen *et al.* 1998). Within this region, the method detected amino acid S662 (groups 3 and 30), N673 (group 31), and I674 (group 32). Not surprisingly, we also detected coevolution between these sites and others involved in interaction with eNOS, AKT, and glucocorticoid receptors (GRs).

The detected coevolution between sites from the N-terminal domain (A112–S115), involved in nucleotide binding, and sites from the middle segment and C-terminal domain (K445 and S446), involved in binding unfolded polypeptides, correlates with their functional relationships (Johnson *et al.* 2000). The coevolution between the N-terminal (S115) and the C-terminal (K636) ATP-binding domains supports previous functional data (Söti *et al.* 2002) and also correlates with the coevolution between the ATP-binding pocket in the N-terminal domain (A112 and D113 in group 17; S115 in group 23) and the ATP modulator domain (E351 in groups 17 and 23). The detected coevolution of sites belonging to distinct binding pockets of accessory elements (*e.g.*, in group 24) supports works pinpointing the temporal regulation and coordinated interaction of cochaperones and Hsp90 (Chen *et al.* 1998). Other sites have been identified as coevolving with DD2 including, remarkably, the site W296 in which mutation in humans interferes in the interaction between Hsp90 and Akt (Meyer *et al.* 2003).

Figure 4.—Coevolution between amino acids in Gag protein from HIV-1. (A) groups of coevolution detected prior to clade removal. Sites detected as having undergone adaptive evolution by Yang *et al.* (2003) are underlined. (B) Subtypes (A–D, F–H, J, and K) are colored differently and the coevolving sites detected in the lineages leading to their ancestors (marked in boldface type on the phylogeny) are plotted on a schematic representation of the 55-kDa precursor (lengths are not exactly proportional to the gene size). Bootstrap values for the ancestral branch of each subtype are also indicated.

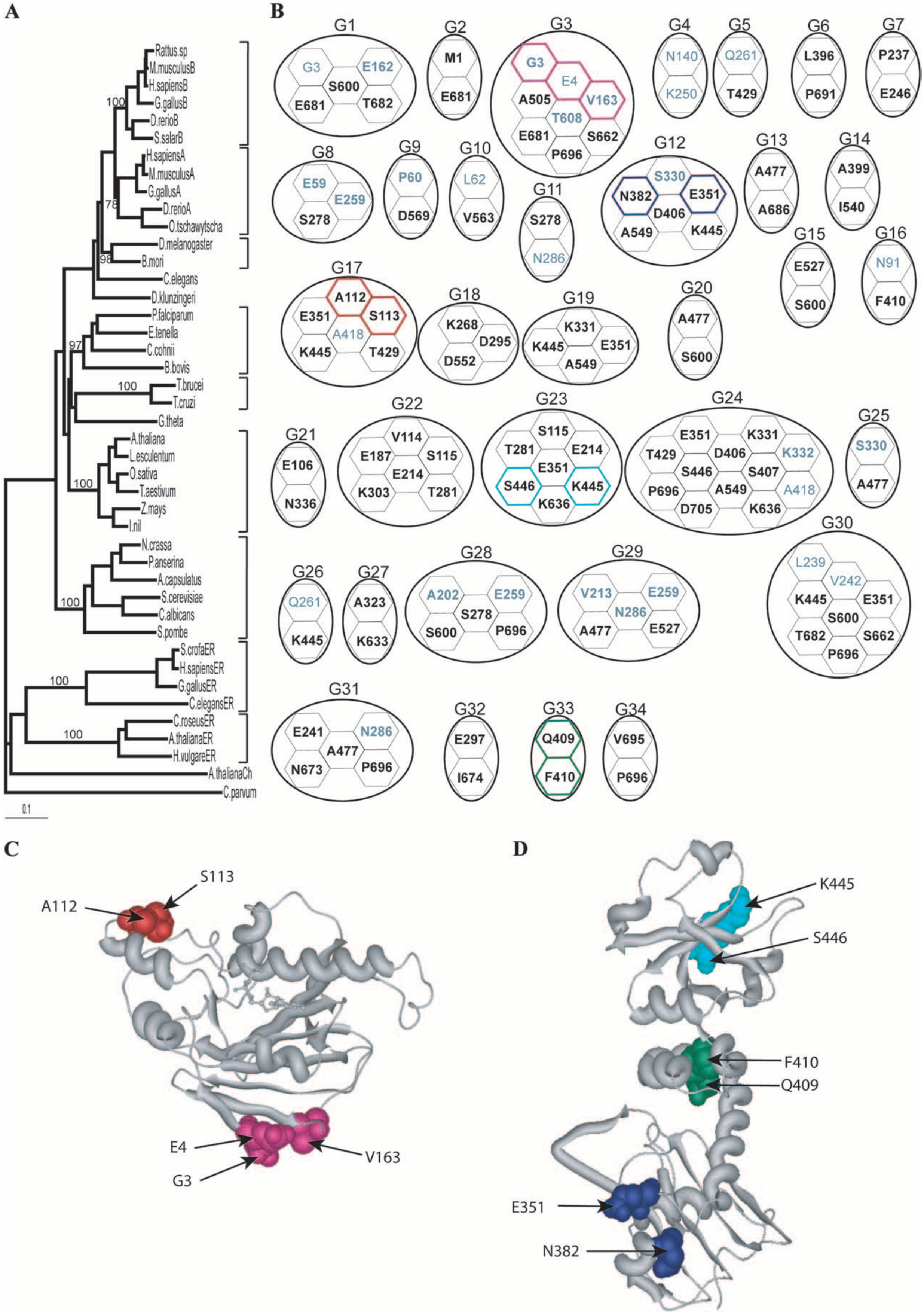

A
Rattus.sp
M.musculusB
H.sapiensB
G.gallusB
D.rerioB
S.salarB
H.sapiensA
M.musculusA
G.gallusA
D.rerioA
O.tschawytscha
D.melanogaster
B.mori
C.elegans
D.klunzingeri
P.falciparum
E.tenella
C.cohnii
B.bovis
T.brucei
T.cruzi
G.theta
A.thaliana
L.esculentum
O.sativa
T.aestivum
Z.mays
I.nil
N.crassa
P.anserina
A.capsulatus
S.cerevisiae
C.albicans
S.pombe
S.crofaER
H.sapiensER
G.gallusER
C.elegansER
C.roseusER
A.thalianaER
H.vulgareER
A.thalianaCh
C.parvum
100
78
98
97
100
100
100
100
100
0.1
B
G1 G3 E162 S600 E681 T682
G2 M1 E681
G3 G3 E4 V163 A505 T608 E681 S662 P696
G4 N140 K250
G5 Q261 T429
G6 L396 P691
G7 P237 E246
G8 E59 E259 S278
G9 P60 D569
G10 L62 V563
G11 S278 N286
G12 S330 N382 E351 D406 A549 K445
G13 A477 A686
G14 A399 I540
G15 E527 S600
G16 N91 F410
G17 A112 E351 S113 A418 K445 T429
G18 K268 D295 D552
G19 K331 K445 E351 A549
G20 A477 S600
G21 E106 N336
G22 V114 E187 S115 E214 K303 T281
G23 S115 T281 E214 E351 S446 K445 K636
G24 E351 K331 T429 D406 K332 S446 S407 P696 A549 A418 D705 K636
G25 S330 A477
G26 Q261 K445
G27 A323 K633
G28 A202 E259 S278 S600 P696
G29 V213 E259 N286 A477 E527
G30 L239 V242 K445 E351 S600 T682 S662 P696
G31 E241 N286 A477 N673 P696
G32 E297 I674
G33 Q409 F410
G34 V695 P696
C
S113
A112
E4
V163
G3
D
K445
S446
F410
Q409
E351
N382

TABLE 1

Coevolution analysis in Hsp90

Group	$\bar{D}_C$[a] ± SE	$\bar{\rho}$[b] ± SE	No. aa	P[c] (close aa)	Functional domains
G1	0.6582 ± 0.2858	0.6932 ± 0.0511	5	10	PI3, DD3
G2	0.3979 ± 0.2032	0.6011 ± 0	2	—	DD3
G3	0.8299 ± 0.727	0.7097 ± 0.0460	8	9	PI3, DD2, DD3
G4	1.8966 ± 0.7638	0.5404 ± 0	2	—	—
G5	2.0849 ± 1.1056	0.8405 ± 0	2	—	AED1
G6	2.9019 ± 0.6877	0.8888 ± 0	2	—	PI2, DD3
G7	1.3166 ± 1.0630	0.7843 ± 0	2	—	Ia, Ib
G8	1.2940 ± 0.328	0.7591 ± 0.0269	3	0	Ic
G9	1.2530 ± 0.439	0.7152 ± 0	2	—	DD1
G10	0.3151 ± 0.069	0.7122 ± 0	2	—	DD1
G11	1.0584 ± 0.1471	0.8476 ± 0	2	100	Ic
G12	0.8842 ± 0.5267	0.7433 ± 0.0669	6	6.67	ATPM1, ATPM2, PI2, PI3
G13	0.7462 ± 0.4987	0.8585 ± 0	2	—	PI3, DD3
G14	0.7616 ± 0.5003	0.8370 ± 0	2	—	PI2, PI3
G15	1.4569 ± 0.5853	0.7925 ± 0	2	—	PI3
G16	2.6162 ± 1.8745	0.6822 ± 0	3	—	—
G17	0.7381 ± 0.3536	0.7310 ± 0.0483	6	26.67	ATP-Nt, ATPM1, AED1, PI3
G18	1.2846 ± 0.1012	0.7223 ± 0	3	—	Ic, PI3
G19	1.2898 ± 0.1966	0.7055 ± 0.0253	4	0	ATPM1, PI3
G20	1.0228 ± 0.0287	0.7730 ± 0	2	—	PI3
G21	0.8212 ± 0.4032	0.7338 ± 0	2	—	—
G22	0.8793 ± 0.3314	0.7162 ± 0.0672	6	20	ATP-Nt
G23	0.9247 ± 0.3060	0.6795 ± 0.0491	7	4.76	ATP-Nt, ATPM1, ATP-Ct, PI3, AED2
G24	0.6914 ± 0.6208	0.7499 ± 0.1046	12	21.21	ATPM1, PI2, PI3, ATP-Ct, AED1, AED2
G25	1.0872 ± 0.1199	0.7155 ± 0	2	0	PI3
G26	1.4233 ± 0.4478	0.7755 ± 0	2	—	PI3
G27	0.5972 ± 0.3336	0.7311 ± 0	2	—	PI1, AED2
G28	1.1494 ± 0.1688	0.7713 ± 0.0124	5	0	Ic, PI3, DD3
G29	1.3453 ± 0.4621	0.8241 ± 0.0124	5	10	PI3
G30	1.0847 ± 0.0309	0.7324 ± 0	8	0	ATPM1, PI3, DD2, DD3
G31	1.0825 ± 0.1131	0.8215 ± 0	5	0	PI3, DD2, DD3
G32	0.531 ± 0.1253	0.9482 ± 0	2	—	DD2
G33	0.07 ± 0.007	0.6250 ± 0	2	100	PI2
G34	0.8997 ± 0.2095	0.7695 ± 0	2	100	DD3

[a] Amino acid site variance.
[b] Mean Pearson correlation coefficient.
[c] Proportion of residues pairs three-dimensionally close.

Finally, other sites with no reported functional importance to date have been detected in this study to coevolve with sites involved in protein interaction, ATP binding, or dimerization. All of these sites are exposed in the protein as predicted by the neural network analyses and by the three-dimensional model of Hsp90. These sites might establish additional interactions between Hsp90 domains and protein clients or may be responsible for the stabilization of the Hsp90 dimer through intermonomer interactions.

GroEL is under strong selective constraints of coevolution: GroEL analysis points to three conclusions: Some of the coevolving sites are spatially close, supporting a functional and maybe structural complicity; the main interdomain coevolution took place between the apical and equatorial domains; and amino acid sites under adaptive evolution have also been identified as coevolving between each other.

Previous works have shown that GroEL has been accumulating advantageous mutations to improve its

Figure 5.—Coevolution between amino acids in Hsp90. (A) Phylogenetic tree with the different clades prespecified indicated and the bootstrap support of their ancestral node shown. (B) Groups of coevolution (G1–G34) as detected by the new method. All the sites included in the same solid-lined circle coevolve between each other. Functionally or structurally important sites are in black, non-important sites are in gray, and selected example sites, plotted in the three-dimensional structure, are colored in red, blue, pink, and green. (C) Three-dimensional structure of the N-terminal domain of Hsp90 (Prodromou *et al.* 1997) showing proximal sites colored. (D) Three-dimensional structure of the middle segment of Hsp90 (Meyer *et al.* 2003) showing proximal sites colored.

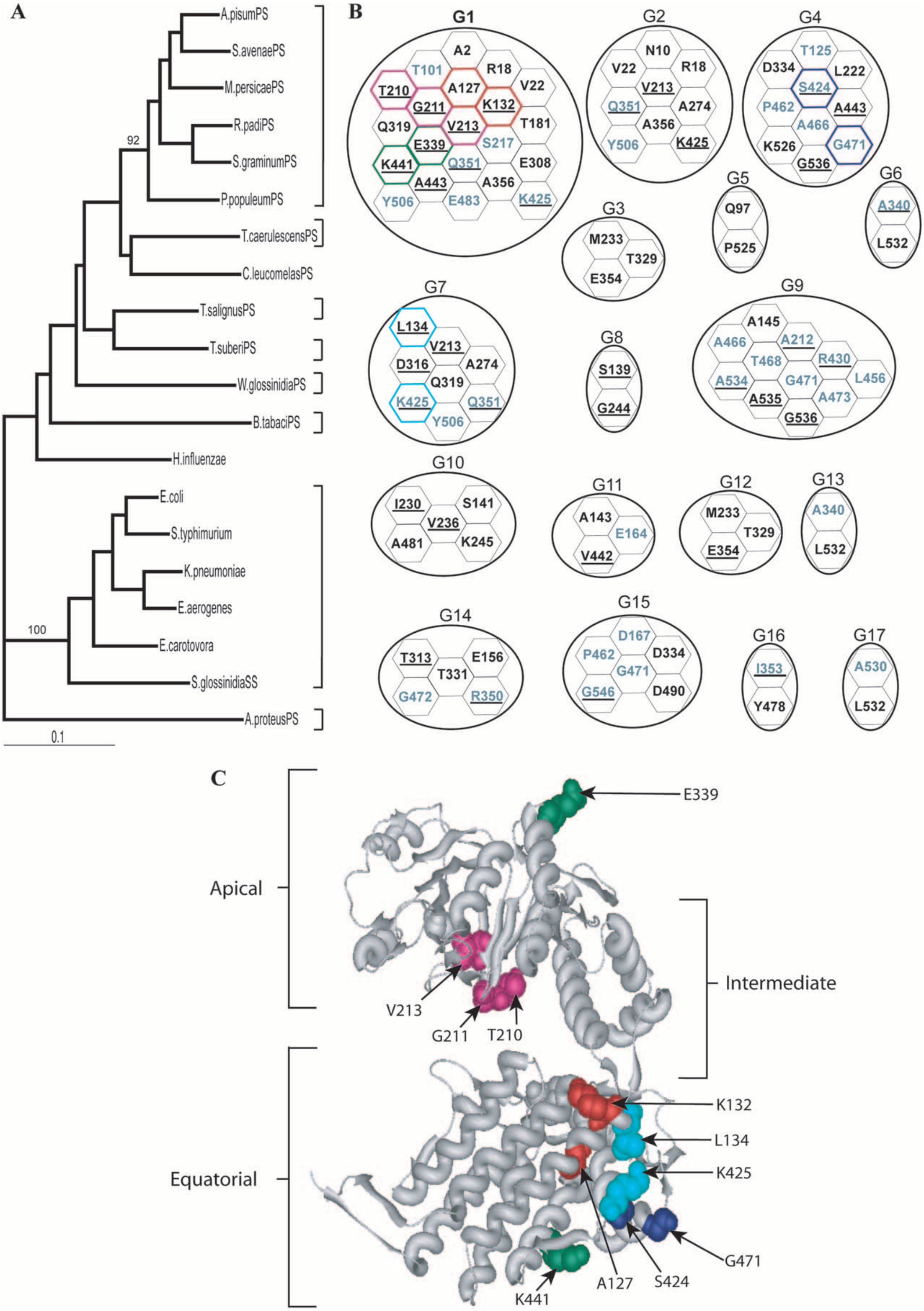

A
A.pisumPS
S.avenaePS
M.persicaePS
R.padiPS
S.graminumPS
P.populeumPS
T.caerulescensPS
C.leucomelasPS
T.salignusPS
T.suberiPS
W.glossinidiaPS
B.tabaciPS
H.influenzae
E.coli
S.typhimurium
K.pneumoniae
E.aerogenes
E.carotovora
S.glossinidiaSS
A.proteusPS
92
100
0.1
B
G1
A2 T101 R18 T210 A127 V22 G211 K132 Q319 V213 T181 E339 S217 K441 Q351 E308 A443 A356 Y506 E483 K425
G2
N10 V22 R18 V213 Q351 A274 A356 Y506 K425
G3
M233 T329 E354
G4
T125 D334 L222 S424 P462 A443 A466 K526 G471 G536
G5
Q97 P525
G6
A340 L532
G7
L134 V213 D316 A274 Q319 K425 Q351 Y506
G8
S139 G244
G9
A145 A466 A212 T468 R430 A534 G471 L456 A535 A473 G536
G10
I230 S141 V236 A481 K245
G11
A143 E164 V442
G12
M233 T329 E354
G13
A340 L532
G14
T313 E156 T331 G472 R350
G15
D167 P462 D334 G471 G546 D490
G16
I353 Y478
G17
A530 L532
C
E339
Apical
Intermediate
V213
G211
T210
K132
L134
K425
Equatorial
G471
A127
S424
K441

TABLE 2

Coevolution analysis in GroEL

Group	$\bar{D}_C$[a] ± SE	$\bar{\rho}$[b] ± SE	No. (aa)	P[c]	Functional domains
G1	0.9937 ± 0.5540	0.6922 ± 0.1417	21	3.33	CER
G2	0.7152 ± 0.3850	0.8031 ± 0.1870	9	5.56	—
G3	1.1824 ± 0.5172	0.8697 ± 0.0118	3	0	SubB, ESB
G4	1.9030 ± 0.9995	0.7813 ± 0.1231	8	10.71	CER
G5	0.7902 ± 0.1665	0.5550 ± 0	2	0	CER
G6	1.7156 ± 0.5226	0.5100 ± 0	2	0	—
G7	0.7781 ± 0.4587	0.8513 ± 0.1053	8	7.14	—
G8	1.2511 ± 0.2326	0.5884 ± 0	2	0	—
G9	1.7240 ± 0.7983	0.7671 ± 0.1435	12	6.06	—
G10	0.9722 ± 0.3063	0.8262 ± 0.1921	5	0	SubB, ATP/Mg-B
G11	0.3915 ± 0.2706	0.7516 ± 0.0633	3	0	—
G12	1.1824 ± 0.5172	0.8697 ± 0.0118	3	0	ESB
G13	1.6982 ± 0.7080	0.6110 ± 0	2	0	—
G14	1.6528 ± 0.7016	0.7666 ± 0.1159	5	0	—
G15	2.4771 ± 0.6675	0.7588 ± 0.1875	4	0	—
G16	2.1029 ± 0.8725	0.9442 ± 0	2	0	—
G17	2.3378 ± 2.1690	0.7335 ± 0	2	0	—

[a] Amino acid site variance.
[b] Mean pearson correlation coefficient.
[c] Proportion of residues pairs three-dimensionally close.

ability to compensate the effect of slightly deleterious and conformational-destabilizing mutations fixed in the proteome of endosymbionts (Fares *et al.* 2002b, 2004). Studies have even suggested that GroEL has been fixing slightly deleterious mutations as a result of the genetic drift operating in the genome of these bacteria (Herbeck *et al.* 2003). The fact that the method has detected amino acids coevolving with spatially close functional sites (Figure 5C) and that these sites underwent adaptive evolution suggests the existence of compensatory changes to maintain the overall structural stability of GroEL. This result can be fully explained under the neutral theory of evolution (Kimura 1983). Under this model, changes altering the amino acid side-chain spatial distribution will disrupt the already optimized performance of the protein's contribution to the organism's fitness. These disadvantageous mutations can be compensated only by changes in spatially proximal sites, where mutations will be fixed by adaptive evolution. The joint contribution of both mutations to the fitness of the organism will be neutral (Fukami-Kobayashi *et al.* 2002).

Advantages and limitations of the method: Many methods developed to detect coevolution (*e.g.*, Göbel *et al.* 1994; Neher 1994) have shown an inability to screen out background correlation, particularly in the presence of phylogenetic relatedness between the sequences, and to distinguish between positive and negative correlation (Pollock and Taylor 1997). As we have shown, CAPS is effective in distinguishing background correlation from true correlations. CAPS analysis can be performed with no knowledge of the phylogenetic relationships among sequences. We have, however, shown that the removal of well-defined clades can serve in the identification of structural/functional coevolution.

Coevolution methods have also been used in an attempt to resolve the docking problem (Göbel *et al.* 1994; Pazos *et al.* 1997). Despite their elegance, most of these methods failed to distinguish unambiguously between coevolution and phylogenetic noise. Correction of BLOSUM values by the sequence divergence and the consideration of functional and structural coevolution in CAPS allows isolation of the true coevolutionary events. It is important to mention that, unlike other studies, atomic distances are not used here as evidence of coevolution but rather as additional supporting information in the identification of functional and structural coevolution.

Figure 6.—Coevolution between amino acids sites in GroEL. (A) Phylogenetic tree used in this study with all the clades or lineages prespecified and the bootstrap value of their ancestral node indicated where appropriate. (B) Groups of coevolution (G1–G17) as detected by the new method. All the sites included in the same solid-lined circle coevolve between each other. Functionally or structurally important sites are in black. Amino acids not detected to be under constraints are in gray. Sites spatially proximal are colored in pink, blue, and red and are shown in the three-dimensional GroEL structure (C) accordingly. Coevolving sites under adaptive evolution (Fares *et al.* 2002b) are underlined. (C) Three-dimensional structure of GroEL (Boisvert *et al.* 1996) showing spatially proximal coevolving sites. Sites in green represent an example of functional coevolution between amino acid sites distant in the structure.

As expected, the method does not lack limitations. For example, saturation of synonymous sites can lead to underestimates of the divergence times, although data sets used in this study did not show such effects. The number of sequences in the alignment also poses a problem when sequences are too divergent, although the sensitivity is improved compared to that of previous methods. Further, constant amino acid sites that are very likely to be functionally important cannot be tested for coevolution using CAPS, although this limitation affects all the methods so far. Moreover, our method assumes that the coevolutionary relationship between a pair of sites remains constant through time. This assumption can be simplistic when analyzing alignments including highly divergent sequences. Nonetheless, the dynamic removal of prespecified phylogenetic clades ameliorates this problem.

Finally, even though coevolutionary analyses can be used to identify protein–protein interaction interfaces, CAPS is not designed for such a purpose. The reason is that, while interaction would necessarily involve coevolution, coevolution does not imply physical interaction. Detecting amino acids involved in protein–protein interactions is a more complex problem, requiring the consideration of other parameters such as solvent accessibility, physiochemical amino acid properties, etc. We have shown here that coevolutionary analyses in biologically key molecules add another dimension to selective constraints analyses and provide more interpretable results.

We are grateful to two anonymous reviewers for important comments and insightful suggestions on the manuscript. This work was supported by Science Foundation Ireland.

LITERATURE CITED

Ané, C., J. G. Burleigh, M. M. McMahon and M. J. Sanderson, 2004 Covarion structure in plastid genome evolution: a new statistical test. Mol. Biol. Evol. **22:** 914–924.

Berezin, C., F. Glaser, J. Rosenberg, I. Paz, T. Pupko *et al.*, 2004 ConSeq: the identification of functionally and structurally important residues in protein sequences. Bioinformatics **20:** 1322–1324.

Berglund, A.-C., B. Wallner, A. Elofsson and D. A. Liberles, 2005 Tertiary windowing to detect positive diversifying selection. J. Mol. Evol. **60:** 499–504.

Boisvert, D. C., J. Wang, Z. Otwinowski, A. L. Horwich and P. B. Sigler, 1996 The 2.4 Å crystal structure of the bacterial chaperonin GroEL complexed with ATP gamma S. Nat. Struct. Biol. **3:** 170–177.

Braig, K., Z. Otwinowski, R. Hegde, D. C. Boisvert, A. Joachimiak *et al.*, 1994 The crystal structure of the bacterial chaperonin *GroEL* at 2.8Å. Nature **371:** 578–586.

Braig, K., P. D. Adams and A. T. Brünger, 1995 Conformational variability in the refined structure of the chaperonin GroEL at 2.8Å resolution. Nat. Struct. Biol. **2:** 1083–1094.

Buchner, J., 1999 Hsp90 & Co.—a holding for folding. Trends Biochem. Sci. **24:** 136–141.

Caplan, A. J., 1999 Hsp90's secrets unfold: new insights from structural and functional studies. Trends Cell Biol. **9:** 262–268.

Chadli, A., I. Bouhouche, W. Sullivan, B. Stensgard, N. McMahon *et al.*, 2000 Dimerization and N-terminal domain proximity underlie the function of the molecular chaperone heat shock protein 90. Proc. Natl. Acad. Sci. USA **97:** 12524–12529.

Chelvanayagam, G., A. Eggenschwiler, L. Knecht, G. H. Connet and S. A. Benner, 1997 An analysis of simultaneous variation in protein structures. Protein Eng. **10:** 307–316.

Chen, S., W. P. Sullivan, D. O. Toft and D. F. Smith, 1998 Differential interactions of p23 and the TPR-containing proteins Hop, Cyp40, FKBP52 and FKBP51 with Hsp90 mutants. Cell Stress Chaperones **3:** 118–129.

Clark, A. G., and T. H. Kao, 1991 Excess nonsynonymous substitution of shared polymorphic sites among self-incompatibility alleles of Solanaceae. Proc. Natl. Acad. Sci. USA **88:** 9823–9827.

Deuerling, E., A. Schulze-Specking, T. Tomoyasu, A. Mogk and B. Bukau, 1999 Trigger factor and DnaK cooperate in folding of newly synthesized proteins. Nature **400:** 693–696.

Dimmic, M. W., and M. J. Hubisz, 2005 Detecting coevolving amino acid sites using Bayesian mutational mapping. Bioinformatics **21:** i126–i135.

Dutheil, J., T. Pupko, A. Jean-Marie and N. Galtier, 2005 A model-based approach for detecting co-evolving positions in a molecule. Mol. Biol. Evol. **22:** 1919–1928.

Fares, M. A., 2004 SWAPSC: sliding-window analysis procedure to detect selective constraints. Bioinformatics **20:** 2867–2868.

Fares, M. A., S. F. Elena, J. Ortiz, A. Moya and E. Barrio, 2002a A sliding window-based method to detect selective constraints in protein-coding genes and its application to RNA viruses. J. Mol. Evol. **55:** 509–521.

Fares, M. A., E. Barrio, B. Sabater-Munoz and A. Moya, 2002b The evolution of the heat-shock protein GroEL from Buchnera, the primary endosymbiont of aphids, is governed by positive selection. Mol. Biol. Evol. **19:** 1162–1170.

Fares, M. A., A. Moya and E. Barrio, 2004 GroEL and the maintenance of bacterial endosymbiosis. Trends Genet. **20:** 413–416.

Fitch, W. M., and E. Markowitz, 1970 An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem. Genet. **4:** 579–593.

Fujita, N., S. Sato, A. Ishida and T. Tsuruo, 2002 Involvement of Hsp90 in signaling and stability of 3-phosphoinositide-dependent kinase-1. J. Biol. Chem. **277:** 10346–10353.

Fukami-Kobayashi, K., D. R. Schreiber and S. A. Benner, 2002 Detecting compensatory covariation signals in protein evolution using reconstructed ancestral sequences. J. Mol. Biol. **319:** 729–743.

Galtier, N., 2004 Sampling properties of the bootstrap support in molecular phylogeny: influence of nonindependence among sites. Syst. Biol. **53:** 38–46.

Gloor, G. B., L. C. Martin, L. M. Wahl and S. D. Dumn, 2005 Mutual information in protein multiple sequence alignment reveals two classes of coevolving positions. Biochemistry **44:** 7156–7165.

Göbel, U., C. Sander, R. Schneider and A. Valencia, 1994 Correlated mutations and residue contacts in proteins. Proteins **18:** 309–317.

Gottlinger, H. G., J. G. Sodroski and W. A. Haseltine, 1989 Role of capsid precursor processing and myristoylation in morphogenesis and infectivity of human immunodeficiency virus type I. Proc. Natl. Acad. Sci. USA **86:** 5781–5785.

Henikoff, S., and J. G. Henikoff, 1992 Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA. **89:** 10915–10919.

Herbeck, J. T., D. J. Funk, P. H. Degnan and J. J. Wernegreen, 2003 A conservative test of genetic drift in the endosymbiotic bacterium Buchnera: slightly deleterious mutations in the chaperonin groEL. Genetics **165:** 1651–1660.

Hughes, A. L., and M. Nei, 1988 Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature **335:** 167–170.

Jeanmougin, F., J. D. Thompson, M. Gouy, D. G. Higgins and T. J. Gibson, 1998 Multiple sequence alignment with Clustal X. Trends Biochem. Sci. **23:** 403–405.

Jones, D. T., W. R. Taylor and J. M. Thornton, 1992 The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. **8:** 275–282.

Johnson, B. D., A. Chadli, S. J. Felts, I. Bounhouche, M. G. Catelli *et al.*, 2000 Hsp90 chaperone activity requires the full-length protein and interaction among its multiple domains. J. Biol. Chem. **275:** 32499–32507.

Kimura, M., 1983 The neutral theory of molecular evolution. Cambridge University Press, Cambridge, UK.

Korber, B. T., R. M. Farber, D. H. Wolpert and A. S. Lapedes, 1993 Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. Proc. Natl. Acad. Sci. USA **8:** 1549–1560.

Landry, S. J., J. Zeilstra-Ryalls, O. Fayet, C. Georgopoulos and L. M. Gierasch, 1993 Characterization of a functionally important mobile domain in GroES. Nature **364:** 255–258.

Lockhart, P. J., M. A. Steel, A. C. Barbrook, D. H. Huson, M. A. Charleston *et al.*, 1998 A covariatoid model explains apparent phylogenetic structure of oxygenic photosynthetic lineages. Mol. Biol. Evol. **15:** 1183–1188.

Lockless, W., and R. Ranganathan, 1999 Evolutionary conserved pathways of energetic connectivity in protein families. Science **286:** 295–299.

Mayer, M. P., and B. Bukau, 1999 Molecular chaperones: the busy life of Hsp90. Curr. Biol. **9:** R322–R325.

Meyer, P., C. Prodromou, B. Hu, C. Vaughan, S. M. Roe *et al.*, 2003 Structural and functional analysis of the middle segment of hsp90: implications for ATP hydrolysis and client protein and cochaperone interactions. Mol. Cell **11:** 647–658.

Neher, E., 1994 How frequent are correlated changes in families of protein sequences? Proc. Natl. Acad. Sci. USA **91:** 98–102.

Pazos, F., M. Helmer-Citterich, G. Ausiello and A. Valencia, 1997 Correlated mutations contain information about protein-protein interaction. J. Mol. Biol. **271:** 511–523.

Pollock, D. D., and W. R. Taylor, 1997 Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution. Protein Eng. **10:** 647–657.

Pollock, D. D., W. R. Taylor and N. Goldman, 1999 Coevolving protein residues: maximum likelihood identification and relationship to structure. J. Mol. Biol. **287:** 187–198.

Pratt, W. B., 1998 The hsp90-based chaperone system: involvement in signal transduction from a variety of hormone and growth factor receptors. Proc. Soc. Exp. Biol. Med. **217:** 420–434.

Pritchard, L., and M. J. Dufton, 2000 Do proteins learn to evolve? The hopfield network as a basis for the understanding of protein evolution. J. Theor. Biol. **202:** 77–86.

Pritchard, L., P. M. O. Bladon, J. J. Mitchell and M. J. Dufton, 2001 Evaluation of a novel method for the identification of coevolving protein residues. Protein Eng. **14:** 549–555.

Prodromou, C., S. M. Roe, R. O'Brien, J. E. Ladbury, P. W. Piper *et al.*, 1997 Identification and structural characterization of the ATP/ADP-binding site in the Hsp90 molecular chaperone. Cell **90:** 65–75.

Prodromou, C., G. Siligardi, R. O'Brien, D. N. Woolfson, L. Regan *et al.*, 1999 Regulation of Hsp90 ATPase activity by the tetratricopeptide repeat (TPR)-domain co-chaperones. EMBO J. **18:** 754–762.

Robertson, D. L., J. P. Anderson, J. A. Bradac, J. K. Carr, B. Foley *et al.*, 2000 HIV-1 nomenclature proposal. Science **288:** 55–56.

Shindyalov, I. N., N. A. Kolchanov and C. Sander, 1994 Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? Protein Eng. **7:** 349–358.

Söti, C., A. Racz and P. Csermely, 2002 A nucleotide-dependent molecular switch controls ATP binding at the C-terminal domain of Hsp90. N-terminal nucleotide binding unmasks a C-terminal binding pocket. J. Biol. Chem. **277:** 7066–7075.

Süel, G. M., S. W. Lockless, A. A. Wall and R. Ranganathan, 2003 Evolutionary conserved networks of residues mediate allosteric communication in proteins. Nat. Struct. Biol. **10:** 59–69.

Suzuki, Y., 2004 Three-dimensional window analysis for detecting positive selection at structural regions of proteins. Mol. Biol. Evol. **21:** 2352–2359.

Suzuki, Y., and T. Gojobori, 1999 A method for detecting positive selection at single amino acid sites. Mol. Biol. Evol. **16:** 1315–1328.

Swofford, D. L., 1998 *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)*. Sinauer Associates, Sunderland, MA.

Taylor, W. R., and K. Hatrick, 1994 Compensating changes in protein multiple sequence alignments. Protein Eng. **7:** 341–348.

Thulasiraman, V., C. F. Yang and J. Frydman, 1999 In vivo newly translated polypeptides are sequestered in a protected folding environment. EMBO J. **18:** 85–95.

Tillier, E. R., and R. A. Collins, 1995 Neighbor-joining and maximum-likelihood with rna sequences—addressing the interdependence of sites. Mol. Biol. Evol. **12:** 7–15.

Tillier, E. R., and T. W. Lui, 2003 Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. Bioinformatics **19:** 750–755.

Travers, S. A., M. J. O'Connell, G. P. McCormack and J. O. McInerney, 2005 Evidence for heterogeneous selective pressures in the evolution of the env gene in different human immunodeficiency virus type 1 subtypes. J. Virol. **79:** 1836–1841.

Tuffley, C., and M. Steel, 1998 Modelling the covarion hypothesis of nucleotide substitution. Math. Biosci. **147:** 63–91.

Wain-Hobson, S., P. Sonigo, O. Danos, S. Cole and M. Alizon, 1985 Nucleotide sequence of the AIDS virus, LAV. Cell **40:** 9–17.

Yang, W., J. P. Bielawski and Z. Yang, 2003 Widespread adaptive evolution in the human immunodeficiency virus type 1 genome. J. Mol. Evol. **57:** 212–221.

Yang, Z., R. Nielsen, N. Goldman and A.-M. K. Pedersen, 2000 Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **153:** 1077–1089.

Communicating editor: S. Yokoyama